CLAIMS

What is claimed is:

1. A method comprising:

creating a suffix tree to determine the frequency of phrases within a text corpus;

specifying a set of frequently occurring phrases; and

filtering the set of frequently occurring phrases to determine a set of entity name and jargon term candidates.

2. The method of claim 1 further comprising:

sorting each phrase of the set of frequently occurring phrases in inverse lexicographical order prior to filtering the set of frequently occurring phrases.

3. The method of claim 1 wherein the text corpus is preprocessed.

4. The method of claim 3 wherein the text corpus is text of a human language.

5. The method of claim 4 wherein the human language is Chinese.

6. The method of claim 4 wherein filtering the set of frequently occurring phrases includes comparing a component word of a phrase to a dictionary of common words and excluding the phrase from the set of entity name and jargon term candidates if the component word is a common word.

7. The method of claim 4 further comprising:

reducing the set of entity name and jargon term candidates by applying natural language processing rules.

8. The method of claim 4 wherein the natural language processing rules are rules selected from the list consisting of morphological rules, semantic rules, and syntactic rules.

9. A machine-readable medium containing instructions which, when executed by a processor, cause the processor to perform a method, the method comprising:

creating a suffix tree to determine the frequency of phrases within a text corpus;

specifying a set of frequently occurring phrases; and

filtering the set of frequently occurring phrases to determine a set of entity name and jargon term candidates.

10. The machine-readable medium of claim 9 wherein the method further comprises: sorting each phrase of the set of frequently occurring phrases in inverse lexicographical order prior to filtering the set of frequently occurring phrases.

11. The machine-readable medium of claim 9 wherein the text corpus is preprocessed.

12. The machine-readable medium of claim 11 wherein the text corpus is text of a human language.

13. The machine-readable medium of claim 12 wherein the human language is Chinese.

14. The machine-readable medium of claim 12 wherein filtering the set of frequently occurring phrases includes comparing a component word of a phrase to a dictionary of common words and excluding the phrase from the set of entity name and jargon term candidates if the component word is a common word.

15. The machine-readable medium of claim 12 wherein the method further comprises:

reducing the set of entity name and jargon term candidates by applying natural language processing rules.

16. The machine-readable medium of claim 12 wherein the natural language processing rules are rules selected from the list consisting of morphological rules, semantic rules, and syntactic rules.

17. A system comprising:

a memory having stored therein executable instructions which when executed by a processor, cause the processor to perform operations comprising:

creating a suffix tree data structure, the suffix tree data structure storing phrase frequency data for a text corpus;

using the phrase frequency data to specify a set of frequently occurring phrases; and

filtering the set of frequently occurring phrases to determine a set of entity name and jargon term candidates; and

a processor to execute the instructions.

18. The system of claim 17 wherein the operations further comprise:

sorting each phrase of the set of frequently occurring phrases in inverse lexicographical order prior to filtering the set of frequently occurring phrases.

19. The system of claim 17 wherein the text corpus is preprocessed.

20. The system of claim 19 wherein the text corpus is text of a human language.

21. The system of claim 20 wherein the human language is Chinese.

22. The system of claim 20 wherein filtering the set of frequently occurring phrases includes comparing a component word of a phrase to a dictionary of common words and excluding the phrase from the set of entity name and jargon term candidates if the component word is a common word.

23. The system of claim 20 further comprising:

reducing the set of entity name and jargon term candidates by applying natural language processing rules.

24. The system of claim 20 wherein the natural language processing rules are rules selected from the list consisting of morphological rules, semantic rules, and syntactic rules.